NN/LM Guidelines for Digitization Awards from NLM

Table of Contents

Introduction

This guidance was developed to help NN/LM Member libraries achieve success in their NLM funded digitization initiatives. These guidelines are designed to provide digital project managers specifications for quality digitization of printed text, manuscripts, photographs, rare books, slides, graphic arts, audio recordings, audiovisuals, maps and more.

There are numerous standards from which to choose for the many aspects of a digitization project. The standards tackle issues such as file format of the end product, metadata (both descriptive and technical), and recommended image quality specifications, providing links to relevant guidance. Format selection depends on the material that is to be scanned (books, microforms, video, or audio) and the intended use of the digital files (preservation masters or access copies). These guidelines are meant to provide recommendations for the best possible outcomes for your digitization projects. The technical specifications in Appendix I include both minimum and optimal recommendations, with the understanding that grantees would at least follow the minimum technical standards to the best of their ability.

Content

The following guidance for selection of content to digitize is based on recommendations from a variety of sources and is non-specific regarding what to digitize. Each institution is encouraged to make reasonable decisions about what to digitize with funds provided through the NN/LM.

Likely candidates for digitization may include the following:

- Collections of historical significance
- Collections that are unique to you, or generally rare, and that have not been already digitized by other institutions and made publicly available. Examples include manuscript collections, photographic collections, and locally produced audiovisuals, in addition to institutional archives.
- Collections directly pertaining to your institution, or geographic region. Examples include annual reports, Board minutes, yearbooks, scrapbooks, photographs, newsletters, and archival materials of many types.
- Materials that are out of copyright.
- Materials that support the celebration of a notable contribution of your institution or mark a major milestone.
- Finally, materials where the prospect of digitization could assist in their acquisition, by meeting the wishes of a donor, for example.

Note: A check of HathiTrust or Internet Archive is recommended for almost any project to see if your proposed content has already been digitized; however, a prior appearance in Google Books should not preclude your digitization decision.

Protection of Materials: Collection materials should be evaluated to determine if they are sufficiently stable for placement on an image capture device. Care should be taken to stabilize and repair fragile items before they are digitized. If there is risk of permanent damage to an item, it should not be scanned.

Copyright

Copyright issues must be taken into consideration before starting any digitization program for materials that are to be publicly displayed on the Web.

The first step is to determine whether the item is protected by copyright or whether it is in the public domain. If the material is protected by copyright, the library will need to obtain permission from the copyright owner before making the digitized copy available. If the item is in the public domain, the library does not need permission to digitize it and make it available.

A useful guide to determining the copyright status of a work, published or not, is "Copyright Term and the Public Domain in the United States," produced by Cornell University and updated annually. It is found at:

http://copyright.cornell.edu/resources/publicdomain.cfm

Other useful resources:

http://www.copyright.gov/circs/circ01.pdf -- Copyright Basics from Library of Congress
http://librarycopyright.net/resources/genie/ -- American Library Association, The Copyright Genie

Privacy

While copyright gives the creator the right to reproduce a work, prepare derivatives, distribute copies, perform and display a work publicly, the rights of privacy are subject to law and moral reasoning.

The right of publicity "prevents the unauthorized commercial use of an individual's name, likeness, or other recognizable aspects of one's persona. It gives an individual the exclusive right to license the use of their identity for commercial promotion." With both the rights of publicity and privacy, common sense will be your best guide in deciding whether digitizing and making content publicly available would place you at risk of violation.

For further information see the Society of American Archivist's: External Ethics, Values and Legal Affairs standards: http://www2.archivists.org/standards/external/93

The Society of American Archivists also offers courses on copyright and privacy for digitization.

Accessibility

The Department of Health and Human Services accessibility requirements, requires that all Federal agencies are obligated to make all electronic and information technology (EIT) that they develop, maintain or use compliant with Section 508 of the Rehabilitation Act.

508 Compliance information is located at http://www.section508.gov.

The National Library of Medicine, a part of the National Institutes of Health, U.S. Department of Health and Human Services; and recipients of National Library of Medicine funding through the NN/LM must also meet these requirements.

Scanning Hardware and Software

The hardware and software used to capture and manage digital images is critical to the success of any digitization project. The most obvious important piece of equipment in a digitization project is a good scanner or digital camera. Listed below are some qualities to look for, and some advice about maintaining your equipment. This document from the Library of Congress offers specific advice about what to look for in a scanner: http://www.loc.gov/rr/print/tp/LookForAScanner.pdf.

Hardware (Minimum Requirements)

Device: a good quality factory-calibrated flatbed or overhead scanner or frame-mounted digital camera(s) with 24-bit color, grayscale and bitonal capture capability in the 300-600 dpi range without distortion or loss of critical features present in the items being digitized.

Calibration: Periodic calibration with standardized target; resolution and focus checks as needed.

Maintenance: In addition to manufacturer's warranty and technical support, an ongoing maintenance agreement, including parts replacement and notification of software upgrades, is recommended.

Hardware (Optimal Requirements)

Device: High-end factory calibrated flatbed or overhead scanner or frame-mounted digital camera(s) with 24-bit color, grayscale and bitonal capture capability in the 300-600 dpi range without distortion or loss of critical features present in the items being digitized; equipped with cradle or supports that enable content capture without risk of damage to material being scanned; has positive ergonomic features.

Calibration: Device installed and tested by supplier using a standardized calibration target system capable of measuring actual resolution. Supplier performs capture system color balancing at time of installation.

Maintenance: Maintenance agreement in place; supplier provides technical support, maintenance and software upgrades, including parts replacement, periodic (minimum annual) on-site check of resolution, focus and color accuracy.

Additional Hardware for Audio Capture:

Minimum: PC soundcard; commercial audio analog-to-digital converter
Optimal: commercial audio analog-to-digital converter

Software:

Minimum: Access-based collection management software (e.g. ContentDM), files backed up per institutional IT policy. OR content sent to external trusted repository.

Optimal: Preservation-minded system (e.g. Fedora, Rosetta) with access software on top, files backed up to remote location OR content sent to external trusted repository.

Metadata

There are multiple categories of metadata: Descriptive, Structural, Administrative, Rights Management and Preservation. For a description of each and a general understanding of metadata see http://www.niso.org/publications/press/UnderstandingMetadata.pdf.

The Library of Congress maintains a number of digital library standards. METS, the Metadata Encoding and Transmission Standard is a schema for encoding descriptive, administrative and structural metadata for digital objects. It includes technical metadata about the hardware and software used to scan the material http://www.loc.gov/standards/mets.

The PREMIS Data Dictionary for Preservation Metadata is the international standard for metadata intended to ensure the long term usability of digital objects http://www.loc.gov/standards/premis/.

Quality Assurance

Digital content must be checked for quality, legibility and completeness. For best results, QA of 100% of the captured images is recommended. Scanned text pages should be de-skewed, cropped, legible and in the correct order. Blurred text or images must be replaced with legible files. Deficiencies in the original printed material should be identified by explanatory text pages (e.g. “missing pages 50-51”, “foldout missing,” “Plate XIV missing” etc.). Pictorial content should not contain unwanted artifacts that alter or distort its appearance. Page-by-page checking is recommended for digitized books. File naming should be consistent for individual works and sets of works.

Audio files must be checked for sound quality, skipped content, noises introduced during digital reproduction and other problems not evident in the recording from which the digital copy was made.

Audiovisual files must be checked for sound and viewing quality including skipped content, drop-outs, visual defects and other problems not evident in the original film or tape.

Embedded metadata and technical and descriptive metadata associated with digital content, must be checked for accuracy and completeness. Updated metadata associated with corrected files should replace the metadata that was generated for incorrect files. Ideally, checksums should be used to confirm that data has not been corrupted whenever files are moved or corrected.

Storage and Access

The following list of digital archives is representative of appropriate places to store and share digitized content in addition to any internal digital archive at your institution:

1. The Medical Heritage Library (MHL) **(http://www.medicalheritage.org/)**

- Digital collaborative of the historical collections of several major health sciences libraries in the United States, Canada, and Great Britain, including the National Library of Medicine
- Over 40,000 digitized books, videos, and audio files relating to the history of health and medicine. MHL does not accept scanned manuscript material or still images at this time.
- MHL uses Internet Archive (see below) to manage and provide access to contributed resources.
- MHL seeks new contributors and is strongly interested in learning about digitization projects using medical historical collections.

2. Internet Archive (http://archive.org)

- Non-profit digital library with extensive holdings of freely available texts, films, sound recordings and archived webpages
- No charge to contribute content, but large numbers of textual resources are most easily contributed via one of Internet Archive's own scanning centers (http://archive.org/scanning/)
- Descriptive metadata for contributed content can be harvested (via OAI-PMH) from Internet Archive for local uses.

3. HathiTrust Digital Library (www.hathitrust.org/)

- Repository of digitized texts with public access to public domain material
- Non-profit partnership of 60+ research libraries
- Contributing content requires membership, with a participation fee based on the overlap of the library's holdings with those held in the HathiTrust.
- Growing number of web services and bulk data downloads of contributed content available. Some may be restricted to participating libraries
- HathiTrust's full-text holdings are discoverable through recent library catalog/discovery products, including those by OCLC, EBSCO and Ex Libris.

4. Flickr Commons (http://www.flickr.com/commons)

- Sub-site of Yahoo's Flickr image hosting service focusing on cultural heritage institutions
- Allows metadata contributions from the public via the Flickr website
- Contributed content must be in the Public Domain
- Yahoo must approve registrations, and participating institutions must agree to Terms of Service

5. YouTube (http://www.youtube.com)

- Google's popular video-sharing service.
- Non-profit institutional channels can be created. See http://www.youtube.com/nonprofits
- Tight integration with Google Search aids discovery of contributed content
- Optional transcription and captioning services, however note that YouTube is not considered to be a Section 508 compliant website per HHS guidance: (http://www.hhs.gov/web/socialmedia/getting_started/youtube_guidance.html#accessibility)

Note: YouTube is not a 508 compliant site. Ideally, videos placed on YouTube should be captioned for 508 compliance. Any videos placed on YouTube created under award by the NN/LM, should also be placed on a .gov site which is compliant.

Appendix I: Scanning Specifications for Format Types

Format Type	Minimum	Optimal
Format: Ink on paper. Printed bound volumes; pamphlets; broadsides, advertising and other graphic illustrations; printed single sheets; maps; flat paper art objects; manuscript pages.	Full color PDF derived from good quality master file; includes OCR for use with assistive technologies (e.g. screen readers). Manuscripts may benefit from bitonal derivative images for readability.	Uncompressed 24-bit color TIFF or JPEG2000 master image of each scanned page, minimum 400 dpi; full color PDF with OCR for use with assistive technologies (e.g. screen readers); structure and sequence of images for books or other collection items is described in XML. Manuscripts may benefit from bitonal derivative images for readability.
Format: Photographic prints, including black-and-white, monochrome and color photographs; legacy photographic formats (e.g. albumen prints). *Note: consult FADGI guidance (pages 62-63) for additional information	Color or grayscale image in a widely used output format (e.g. high quality JPEG derived from a high quality master file).	Uncompressed color or grayscale TIFF or JPEG2000 scaled to conform to the dimensions of the original photographic print.
Format: Photographic films, including black-and-white negatives and color camera originals. *Note: consult FADGI guidance (pages 60-61) for additional information	Color or grayscale image in a widely used output format (e.g. high quality JPEG derived from a high quality master file).*	Uncompressed color or grayscale TIFF or JPEG2000 scaled to conform to the dimensions of the original photographic print.* Note: Because of technical issues involved in transmission scanning, it is recommended that films be scanned by a qualified professional scanning service.

Format Type	Minimum	Optimal
Format : Audio **Note: consult Indiana/Harvard project Sound Directions for additional information**	Recording output to CD-quality WAV, or Broadcast WAV (44.1khz/16bit) master, MP3 access derivative , transcript	Recording output to higher bitrate WAV, or Broadcast WAV (48 or 96Khz/24bit) master, MP3 access derivative, transcript
Format: Audiovisuals *Note: consult FADGI guidance for additional information	Video written to MPEG2 or MPEG4, transcript.	Video written to an uncompressed format (e.g. Motion JP2000), transcript and caption file.

Appendix II: Selected Additional Digitization Resources

1. **Federal Agencies Digitization Guidelines Initiative** (FADGI), http://www.digitizationguidelines.gov/, a collaborative effort by federal agencies formed in 2007 to define common guidelines, methods, and practices to digitize historical content in a sustainable manner.

   FADGI maintains two working groups:
   - Still Images Working Group, which includes textual as well as graphical content
   - Audio-Visual Working Group, covering sound, video recordings and motion picture film

2. **National Archives and Records Administration (NARA)** ***Reformatting Approaches***
http://www.archives.gov/preservation/products/definitions/reformatting.html

- Divides outcomes into maximum, median and distribution products (e.g. "preservation master," "reproduction master," "distribution copy.")
- To view specific requirements, select the desired product (e.g. "Textual Maximum – Color") and click open to view recommended file properties, quality control actions, rationale and other information.
- Presents use cases for digitization products based on project scope and desired outcomes

3. **Joint Information Systems Committee (JISC)**
Basic Guidelines for Image Capture and Optimisation.
http://www.jiscdigitalmedia.ac.uk/stillimages/advice/basic-guidelines-for-image-capture-and-optimisation Useful checklist for digitization projects.
Choosing a File Format for Digital Still Images.
http://www.jiscdigitalmedia.ac.uk/stillimages/advice/choosing-a-file-format-for-digital-still-images Basic format selection guidance.

4. **Northeast Document Conservation Center**, another rich source of high-quality information. http://www.nedcc.org/resources/leaflets.list.php